D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions

Sammy Christen[1] Muhammed Kocabas[1,2] Emre Aksan[1]
Jemin Hwangbo[3] Jie Song[1†] Otmar Hilliges[1]
[1]Department of Computer Science, ETH Zurich [2]Max Planck Institute for Intelligent Systems, Tübingen
[3] Department of Mechanical Engineering, KAIST

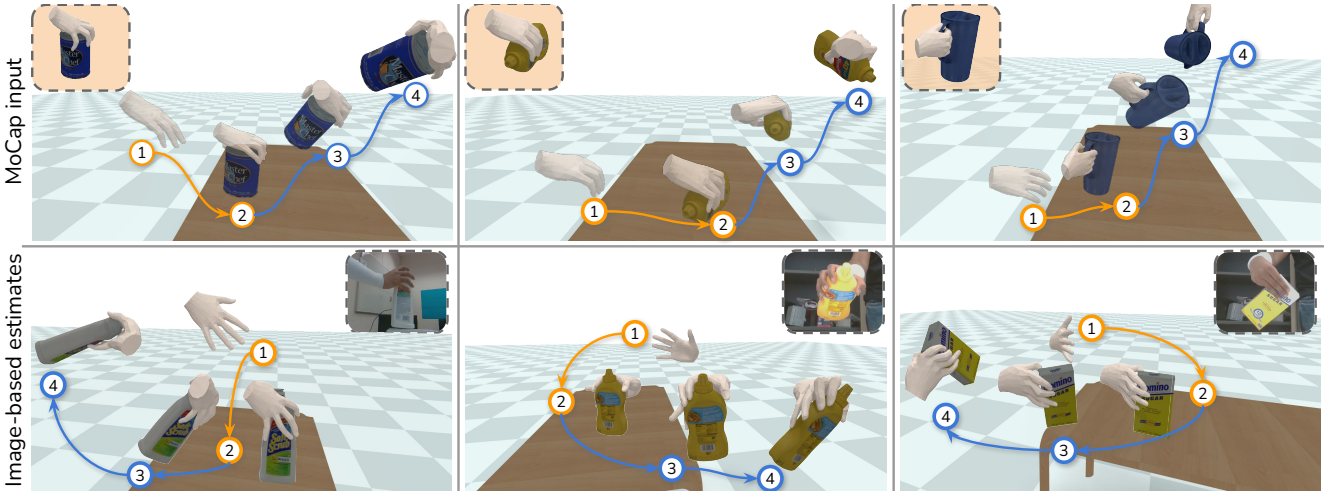

Figure 1. **Dynamic Grasp Synthesis**: Our method learns diverse grasps from static grasp labels (shown in insets), originating from existing datasets, grasp synthesis or image-based estimates. Our approach can then synthesize diverse dynamic sequences with the objects in-hand. We decompose the task into: stable grasping ①-②, followed by the synthesis of a 3D global motion to move the object into a 6D target pose ③-④. The hand-pose is continuously adjusted to ensure a stable grasp, leading to physically plausible and human-like sequences.

Abstract

We introduce the dynamic grasp synthesis task: given an object with a known 6D pose and a grasp reference, our goal is to generate motions that move the object to a target 6D pose. This is challenging, because it requires reasoning about the complex articulation of the human hand and the intricate physical interaction with the object. We propose a novel method that frames this problem in the reinforcement learning framework and leverages a physics simulation, both to learn and to evaluate such dynamic interactions. A hierarchical approach decomposes the task into low-level grasping and high-level motion synthesis. It can be used to generate novel hand sequences that approach, grasp, and move an object to a desired location, while retaining human-likeness. We show that our approach leads to stable grasps and generates a wide range of motions. Furthermore, even imperfect labels can be corrected by our method to generate dynamic interaction sequences. Video and code are available at: https://eth-ait.github.io/d-grasp/.

1. Introduction

A key problem in computer vision is to understand how humans interact with their surroundings. Because hands are our primary means of manipulation with the physical world, there has been an intense interest in hand-object pose estimation [5, 13–15, 19, 38, 39] and the *synthesis* of *static* grasps for a given object [19, 21, 24, 38]. However, human grasping is not limited to a single time instance, but involves a continuous interaction with objects in order to *move* them. It requires maintaining a stable grasp throughout the interaction, introducing intricate dynamics to the task. This involves reasoning about the complex physical interactions between the dexterous hand and the manipulated object, including collisions, friction, and dynamics. A generative model that can synthesize realistic and physically plausible object manipulation sequences would have many downstream applications in AR/VR, robotics and HCI.

†Corresponding author

We propose the new task of *dynamic grasp synthesis*. Given an object with a known 6D pose and a static grasp reference, our goal is to generate a grasping motion and to move the object to a target 6D pose in a natural and physically-plausible way. This new setting adds several challenges. First, the object geometry and the spatial configuration of the object and the hand need to be considered in continuous interaction. Second, contacts between the hand and object are crucial in maintaining stability of the grasps, where even a small error in hand pose may lead to an object slipping. Moreover, contact is typically unobservable in images [10] and measuring the stability of a grasp is very challenging in a static setting. Finally, synthesizing sequences of hand motion requires the generation of smooth and plausible trajectories. While prior work investigates the control of dexterous hands by learning from full demonstration trajectories [11, 32], we address the generation of hand motion from only a single-frame grasp reference. This is a more challenging setting, because the generation of human-like hand-object interaction trajectories without dense supervision is not straightforward.

Taking a step towards this goal, we propose *D-Grasp*, which generates physically plausible grasping motions with only a single grasp reference as input (Fig. 1). Concretely, we formulate the *dynamic grasp synthesis* task as a reinforcement learning (RL) problem and propose a policy learning approach that leverages a physics simulation. Our RL-based approach considers the underlying physical phenomena and compensates data scarcity via exploration in the physics simulation. This ensures physical plausibility, e.g., there is no hand-object interpenetration and the fingers exert enough force on the object to hold it without slipping.

Specifically, we introduce a hierarchical framework that consists of a low-level grasping policy and a high-level motion synthesis module. The grasping policy's purpose is to establish and maintain a stable grasp, whereas the motion synthesis module generates a motion to move the object to a user-specified target position. To guide the low-level grasping policy, we require a single grasp label corresponding to a static hand pose, which can be obtained either from a hand-grasping dataset [5, 13], a state-of-the-art grasp synthesis method [19] or via an image-based pose estimator [12]. Crucially, we propose a reward function that is parameterized by the grasp label to incentivize the fingers to reach contact points on the object, leading to human-like grasps. Our high-level motion synthesis module generates motions that move the hand and object to the final target pose. Importantly, the low-level policy continually controls the grasp to not drop the object.

In our experiments, we first demonstrate that samples from motion capture, static grasp synthesis or image-based pose estimates often do not lead to stable grasps when evaluated in a physics simulation (Fig. 4). We then present how our method can learn to produce physically plausible and stable grasps when guided by such labels. Next, we set out to generate motions with the object in-hand to reach a wide range of target poses. We provide an extensive ablation, revealing the importance of the hierarchical approach and the reward formulation for dynamic grasp synthesis.

Our contributions can be summarized as follows: i) We introduce the new task of *dynamic grasp synthesis*. ii) We propose *D-Grasp*, an RL-based method to synthesize physically-plausible and natural hand-object interactions. iii) We show that our method can generate grasp motions with static grasp references, which can originate from motion capture, static grasp synthesis or image-based pose estimation . We will release our code for research purposes.

2. Related Work

Human Grasp Prediction Recently, hand-object interaction has received much research attention. This growth is accelerated by the introduction of datasets that contain both hand and object annotations [1, 2, 5, 8, 13, 23, 38]. Leveraging this data, a large number of methods attempt to estimate grasp parameters, such as the hand and object pose, directly from RGB images [4, 9, 14, 15, 22, 25, 39, 40]. Some predict the mesh of the hand and the object directly [15], or assume a known object and predict its 6DoF in addition to the hand [4, 14, 25, 40]. Others predict 3D keypoints and 6 DoF pose of the object [9, 39] or produce an implicit surface representation of the grasping hands [22]. To improve the prediction accuracy of the grasp, many of these works incorporate additional contact losses [15, 22] or propose a contact-aware refinement step [4, 40]. More directly related are methods that attempt to generate static grasps given an object and sometimes also information about the hand [1, 2, 19, 21, 22, 38, 44]. Generally, these approaches either predict a contact map on the object [1, 2, 19] or synthesize the joint-angle configuration of the grasping hand [21, 22, 38, 44]. [19] propose a hybrid method, where predicted contact maps on objects are used to refine an initial grasp prediction. Some methods have combined these two directions, for example by leveraging contact information to post-process noisy hand pose predictions [12]. [43] generate local grasp motions, given the global motion of the hand and object. Similarly, [41] synthesize hand grasps given full-body and object motions. In summary, all of these works focus on generating static grasps and are purely data-driven. In our work, however, we take into consideration the dynamic nature of human-object interaction and consider the physical plausibility of dynamic grasp-based hand-object interactions by leveraging a physics-driven simulation.

Dexterous Hand Control Different approaches have been used for controlling dexterous hands. Learning-based methods most often resort to an anchored hand for in-hand manipulation tasks [6, 17, 30], which removes the complex-

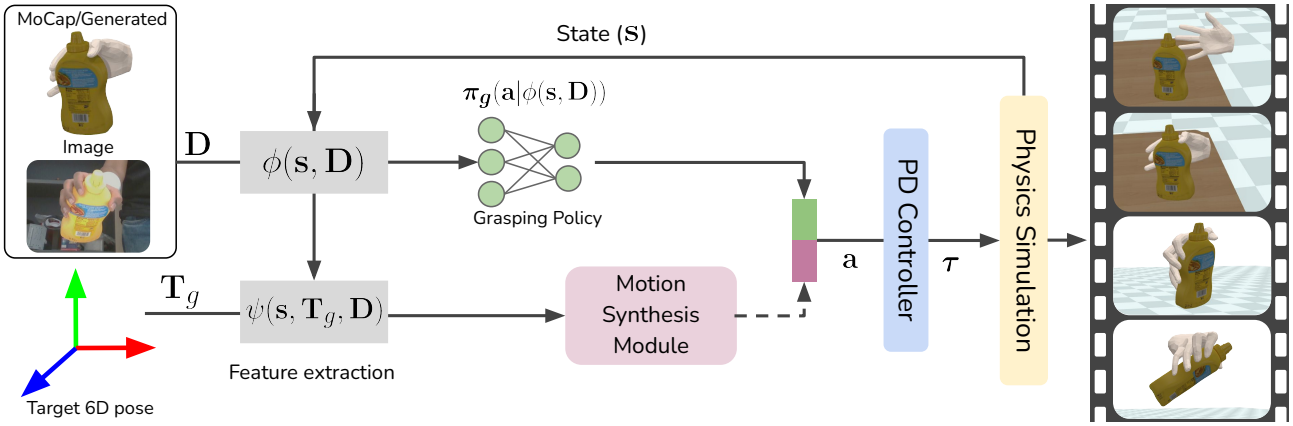

Figure 2. **Method Overview:** Taking a single, static grasp label $\mathbf{D}$ and a target object 6D pose $\mathbf{T}_g$ as input (leftmost), *D-Grasp* produces sequences of dynamic hand-object interactions (rightmost). To do so, we propose a hierarchical framework that consists of a low-level grasping policy $\boldsymbol{\pi}_g(\cdot)$ and a high-level motion synthesis module. In the *grasping* phase, only the grasping policy is active and finds a stable grasp on the object. In the subsequent *motion synthesis* phase, both the grasping policy and the motion synthesis module act concurrently. The actions $\mathbf{a}$ consist of joint targets. These are combined and passed to a PD-controller that computes the required torques $\boldsymbol{\tau}$ to control a MANO-based hand model in a physics simulation. The physics simulation updates the state $\mathbf{s}$ which serves as input to a reward formulation (Section 3.2.2) that forms our supervision signal and incentivizes the hand to approach and grasp the object and to move it to the target 6D pose. We introduce two feature extraction layers ($\phi(\cdot)$ an $\psi(\cdot)$) that utilize the environment state $\mathbf{s}$ and grasp label $\mathbf{D}$ to find a suitable representation for the grasping policy and the motion synthesis module.

ity of generating collision-free trajectories, or rely on expert demonstrations [7, 11, 16, 31, 32], which can be costly to obtain. [32] collect expert trajectories via teleoperation, which they leverage in an RL setup to learn complex manipulation tasks. [11] obtain noisy expert demonstrations from videos and use residual RL to correct the inputs for hand-object interaction tasks. In contrast, we only require a single frame grasp label per sequence. Similar to our work, [7] use a parameterized reward function from single data labels for human-robot interactions, but assume a fixed hand to interact with. [20] propose a modular human manipulation framework, but focus on learning power-grasps for picking up objects. [28] intrinsically motivate a policy to grasp in the affordance region of objects. However, since the policy is only incentivized to grasp in a certain region, the fingers often end up in unnatural configurations. In their follow-up work [27], the authors address this issue by formulating a reward based on hand-object interaction videos. However, the focus is on a single "consensus" grasp reference per object. In our work, we propose a method that learns natural object interactions and generates a wider variety of grasps by explicitly conditioning on the desired contact points and hand pose.

Physics-aware Inference Several recent works have introduced physical awareness to improve purely data-driven approaches [10,26,29,33,36,37,42]. [29] use a physics simulation to validate the plausibility of a generative model for objects via a stability measure. [10] learn to reason about contacts and forces in hand-object interaction videos by leveraging a physics simulation for supervision. To improve the task of human-pose reconstruction from videos, different methods have added physics-based modules to correct the output of a human-pose estimation model. This is achieved either in a post-processing optimization framework [33,37], with an approxmation of physics [36], or via a reinforcement learning policy that directly corrects the pose estimate [42]. [26] regulate a data-driven policy for egocentric pose estimation with a physics-based policy. They include full-body interactions with larger objects, such as pushing a box. In contrast to these works, we introduce the novel task of dynamic hand-object interactions, which involves more fine-grained control of the dexterous human hand and has to adhere to the dynamics and displacement of the object of interest. The task also introduces additional complexities due to the increased amount of collision detection queries required for accurately modeling the contacts. To the best of our knowledge, ours is the first method that studies this task and constitutes an important first step into an important direction for human-object interaction.

3. Method

We propose *D-Grasp*, an RL-based approach that leverages a physics simulation for the *dynamic grasp synthesis* task (Fig. 2). Our model requires a static grasp label consisting of the hand's 6D global pose and local pose for the fingers. We split the task into two distinct phases, namely a *grasping* and a *motion synthesis* phase. In the *grasping* phase, the hand needs to approach an object and find a physically-plausible and stable grasp. In the *motion synthe-*

sis phase, the hand has to bring the object into the 6D target pose while the grasping policy retains a stable grasp on the object. Therefore, the grasping policy and motion synthesis module act concurrently in this phase. To this end, we follow a hierarchical framework that functionally separates the grasping from the motion synthesis.

In the next section, we define the task setting and provide background on RL and the physics simulation. Thereafter, we present both the *grasping* and *motion synthesis* phases of our method in Sections 3.2 and 3.3, respectively.

3.1. Task Setting

In the *dynamic grasp synthesis* task, we are given a 6D global pose $\mathbf{T}_h$ and 3D local pose $\mathbf{q}_h$ of a hand, and an object pose $\mathbf{T}_o$, where the 6D poses consist of a rotation and translation component $\mathbf{T} = [\mathbf{q}|\mathbf{t}]$. Given a label of a static grasp $\mathbf{D} = (\overline{\mathbf{q}}_h, \overline{\mathbf{T}}_h, \overline{\mathbf{T}}_o)$, the goal is to grasp the object and move it into a 6D goal pose $\mathbf{T}_g$. The grasp label consists of the 6D global pose of the hand $\overline{\mathbf{T}}_h$ and object $\overline{\mathbf{T}}_o$, as well as the target hand pose $\overline{\mathbf{q}}_h$ at the instance of the static grasp.

Simulation Setup To approximate a human-like hand in the physics engine, we create a controllable hand model and integrate information obtained from a statistical parametric hand model (i.e., MANO [34]). We extract the skeleton of the hand to get the relative joint positions and add joint actuators for the control of the hand. Finally, we restrict the joints to be within reasonable limits. In our implementation, we use a unified hand model corresponding to the mean MANO shape. Objects are modeled via meshes from the respective datasets [5, 13]. To further speed up the physics simulation, we approximate simple objects with primitive shapes via mesh alignment during training (e.g., a soup can is approximated by a cylinder). For more complex shapes, we use mesh decimation to reduce the number of vertices. For further details, please refer to supp. material.

Reinforcement Learning We follow the standard formulation of a Markov Decision Process (MDP). The MDP is defined as a tuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, , \mathcal{R}, \gamma, \mathcal{T}, \rho_0, \}$, where $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces, respectively. $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function, $\gamma \in [0, 1]$ a discount factor, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ the deterministic transition function of the environment and $\rho_0 = p(\mathbf{s}_0)$ the initial state distribution. We aim to find a probabilistic policy $\boldsymbol{\pi}(\mathbf{a}_t|\mathbf{s}_t)$ with $\mathbf{a}_t \in \mathcal{A}$ and $\mathbf{s}_t \in \mathcal{S}$, maximizing the expected return $\mathbb{E}_{\mathbf{a}_t \sim \boldsymbol{\pi}(\cdot|\mathbf{s}_t), \mathbf{s}_0 \sim \rho_0} \left[\sum_{i=0}^{T} \gamma^i \mathcal{R}(\mathbf{s}_t, \mathbf{a}_t) \right]$ with $\mathbf{s}_{t+1} = \mathcal{T}(\mathbf{s}_t, \mathbf{a}_t)$ at each timestep t.

State Space The state $\mathbf{s} = (\mathbf{q}_h, \dot{\mathbf{q}}_h, \mathbf{f}, \mathbf{T}_h, \dot{\mathbf{T}}_h, \mathbf{T}_o, \dot{\mathbf{T}}_o)$ entails proprioceptive information about the hand pose in the form of joint angles $\mathbf{q}_h$ and joint angular velocities $\dot{\mathbf{q}}_h$, the forces between the hand and object $\mathbf{f}$, the 6D pose of the wrist $\mathbf{T}_h$ and the global 6D pose of the object $\mathbf{T}_o$ with their corresponding velocities $\dot{\mathbf{T}}_h$ and $\dot{\mathbf{T}}_o$. States are expressed with respect to a fixed global coordinate frame. We show experimentally that learning from the full state space can impede learning over several different grasp labels (Section 4.5). We therefore propose a representation that enables learning of the task in Section 3.2.1.

Action Space We define an action space to control the hand in the physics simulation. The fingers are controlled via one actuator per joint for a total of 45 actuators, to which we add 6 DoF to control the global pose. We employ PD-controllers that take reference joint angles $\mathbf{q}_{\text{ref}}$ as input and compute the torques that should be applied to the joints:

$$\boldsymbol{\tau} = k_p(\mathbf{q}_{\text{ref}} - \mathbf{q}) + k_d\dot{\mathbf{q}} \tag{1}$$

$$\mathbf{q}_{\text{ref}} = \mathbf{q}_b + \mathbf{a}. \tag{2}$$

The policy $\boldsymbol{\pi}$ outputs actions $\mathbf{a}$, which are residual actions that change a bias term $\mathbf{q}_b$. For the finger joints, the bias term is equivalent to the current joint configuration $\mathbf{q}_b = \mathbf{q}_h$. We found this formulation to lead to smoother finger motion and therefore more stable grasps compared to the policy directly predicting $\mathbf{q}_{\text{ref}}$. Note that for simplicity's sake, we use the notation $\mathbf{q}_b$ for all joints, although the first three DoF are translational joints.

3.2. Physically Plausible Grasping

Here we discuss the *grasping* phase. The goal is to approach an object and find a physically plausible grasp. A careful design of the model's input representation is key to learning a successful model for hand-object interactions [43], which we show in our ablations (Section 4.5). Therefore, we introduce a feature extraction layer that converts the information from the physics simulation and grasp label into a suitable representation for model learning.

3.2.1 Feature Extraction for Grasping

Rather than directly conditioning the policy on the state, we apply a feature extraction layer $\phi(\mathbf{s}, \mathbf{D})$ that takes the state and grasp label as input. For consistency, we can reformulate the policy as $\boldsymbol{\pi}_g(\mathbf{a}|\phi(\mathbf{s}, \mathbf{D}))$ (Fig. 2). The function $\phi(\cdot)$ processes information from the grasp label, and applies coordinate frame transformations to achieve invariance w.r.t. global coordinates by transforming it to object-relative coordinates. To this end, the feature extraction layer receives the state $\mathbf{s} = (\mathbf{q}_h, \dot{\mathbf{q}}_h, \mathbf{f}, \mathbf{T}_h, \dot{\mathbf{T}}_h, \mathbf{T}_o, \dot{\mathbf{T}}_o)$ and grasp label $\mathbf{D} = (\overline{\mathbf{q}}_h, \overline{\mathbf{T}}_h, \overline{\mathbf{T}}_o)$ as input. Its output is defined as:

$$\phi(\mathbf{s}, \mathbf{D}) = (\mathbf{q}_h, \dot{\mathbf{q}}_h, \mathbf{f}, \widetilde{\mathbf{T}}_h, \widetilde{\mathbf{T}}_o, \dot{\widetilde{\mathbf{T}}}_o, \dot{\widetilde{\mathbf{T}}}_h, \widetilde{\mathbf{x}}_o, \widetilde{\mathbf{x}}_z, \mathbf{G}). \tag{3}$$

The terms $\mathbf{q}_h$ and $\dot{\mathbf{q}}_h$ are the local joint angles and velocities, whereas $\mathbf{f}$ represents contact force information. The

remaining components are expressed in the wrist's reference frame (denoted by $\widetilde{\cdot}$): the object's 6D pose $\widetilde{\mathbf{T}}_o$ and its linear and angular velocities $\dot{\widetilde{\mathbf{T}}}_o$, the hand's 6D pose $\widetilde{\mathbf{T}}_h$ (relative to the initial wrist pose) and its linear and angular velocity $\dot{\widetilde{\mathbf{T}}}_h$, and the displacement of the object from its initial position $\widetilde{\mathbf{x}}_o$. Furthermore, $\widetilde{\mathbf{x}}_z$ introduces awareness of the vertical distance to the surface where the object rests. Lastly, we include the goal components $\mathbf{G} = [\widetilde{\mathbf{g}}_x|\widetilde{\mathbf{g}}_q|\mathbf{g}_c]$, which incentivize the model to reach contact points on the object. We show that these goal components are crucial for achieving stable grasps in Section 4.5. More specifically, the term $\widetilde{\mathbf{g}}_x$ measures the 3D distance between the current and the target 3D positions (Fig. 3), $\mathbf{x}$ and $\overline{\mathbf{x}}$, respectively. Here, all joints and the fingertips are in the wrist's coordinate frame. Importantly, we compute object-relative target positions from the label $\mathbf{D}$ in order to be invariant to the object 6D pose during the grasping phase.

The term $\widetilde{\mathbf{g}}_q$ represents the angular distance between the current rotations $\mathbf{q}_h$ and target rotations $\overline{\mathbf{q}}_h$ for the joints and the wrist. Finally, $\mathbf{g}_c$ includes the target contact vector $\mathbf{g}_c$, i.e., which finger joints should be in contact with the object. A more detailed description about how we extract target contacts, the applied reference frame conversions, and the coordinate representation for individual components of the state or goal space is provided in supp. material.

3.2.2 Reward Function for Grasping

To incentivize the policy to learn the desired behavior, we need to define a reward function. In our method, we formulate it as follows:

$$r = w_x r_x + w_q r_q + w_c r_c + w_{\text{reg}} r_{\text{reg}}. \tag{4}$$

It comprises a combination between position, angle, contact and regularization terms, respectively. We weigh the reward components with the factors $w_x, w_q, w_c, w_{\text{reg}}$.
The position reward r_x measures the weighted sum of distances between the target $\overline{\mathbf{x}}$ and the current 3D positions $\mathbf{x}$ for every joint (including the wrist):

$$r_x = \sum_{j=1}^{J} w_{x,j} \|\overline{\mathbf{x}}_j - \mathbf{x}_j\|^2. \tag{5}$$

Similarly, the pose reward r_q measures the distance between the current pose and the corresponding target pose in Euler angles and corresponds to the L2-norm of the feature $\widetilde{\mathbf{g}}_q$:

$$r_q = \|\widetilde{\mathbf{g}}_q\|, \tag{6}$$

The contact reward r_c is extracted from the finger parts that should be in contact with the object. Specifically, it is computed as the sum of two terms. The first one represents the fraction of target contacts that the agent has achieved. The second term rewards the amount of force exerted on desired contact points, capped by a factor proportional to the object's weight m_o through a factor λ:

$$r_c = \frac{\widetilde{\mathbf{g}}_c^\top \mathbf{I}_{\mathbf{f}>0}}{\widetilde{\mathbf{g}}_c^\top \widetilde{\mathbf{g}}_c} + \min(\widetilde{\mathbf{g}}_c^\top \mathbf{f}, \lambda m_o). \tag{7}$$

Finally, the reward r^{reg} involves regularization terms on the hand's and object's linear and angular velocities:

$$r_{\text{reg}} = w_{\text{reg},h} \|\dot{\widetilde{\mathbf{T}}}_h\|^2 + w_{\text{reg},o} \|\dot{\widetilde{\mathbf{T}}}_o\|^2. \tag{8}$$

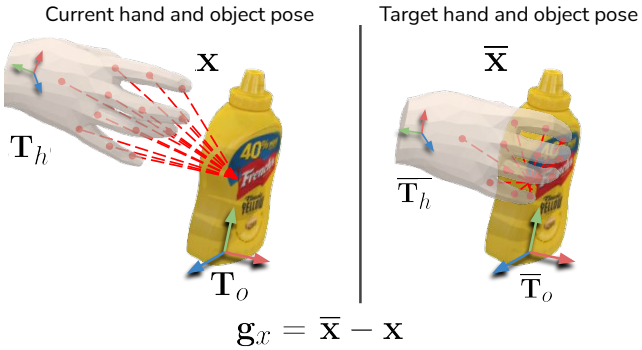

Figure 3. **Target Distance Component $\mathbf{g}_x$.** It incentivizes the policy to reach target points close to the grasp reference label $\mathbf{D}$. We extract the object-relative target 3D joint positions $\overline{\mathbf{x}}$ from $\mathbf{D}$ and compute the distance between $\overline{\mathbf{x}}$ and the current 3D joint positions $\mathbf{x}$ relative to the object's origin. We then convert $\mathbf{g}_x$ into wrist-relative coordinates $\widetilde{\mathbf{g}}_x$.

3.2.3 Wrist-Guidance Technique

To control the global pose during the grasping phase, we introduce a simple but effective technique which we call *wrist-guidance*. Intuitively, we bias the hand to approach the object. To achieve this, we leverage the object-relative target pose, of the hand on the object, obtained from the grasp label $\mathbf{D}$. We then use it as a bias term in the PD-controller of the global 3DoF position. In other words, we set the bias term of the first 3DoF (the translational joints) to $\mathbf{q}_b = \overline{\mathbf{x}}_h$ (Section 3.1), where $\overline{\mathbf{x}}_h$ is the target position which we extract from the label. We find that this technique leads to better performance and faster convergence than using the previous joint positions as bias (Section 3.1), which we show in ablations in Section 4.5.

3.3. Motion Synthesis

We now introduce the motion synthesis module, which is responsible for moving the object from an initial 6D pose into a target 6D pose. It controls only the movement of the wrist, i.e., the first 6DoF of the controllable hand model. In this phase, both the grasping policy described in Section 3.2 and the motion synthesis module are executed concurrently.

		Models	Training set			Test set		
			Success↑	SimDist [mm/s]↓	Interp. [cm^3]↓	Success↑	SimDist [mm/s]↓	Interp. [cm^3]↓
DexYCB	MC	GT+PD	0.31	13.4 ± 9.2	4.59	0.35	13.1 ± 9.1	4.41
		GT+IK	0.39	11.8 ± 9.4	9.23	0.50	9.1 ± 8.5	9.74
		Ours	**0.70**	$\mathbf{5.8 \pm 7.4}$	**1.75**	**0.63**	$\mathbf{8.0 \pm 8.1}$	**1.77**
	SYN	Jiang *et. al* [19]+PD	0.25	12.4 ± 6.4	4.92	0.24	12.7 ± 6.5	4.94
		Ours	**0.75**	$\mathbf{3.9 \pm 7.2}$	**2.84**	**0.73**	$\mathbf{4.6 \pm 6.7}$	**2.81**
HO3D	SYN	Jiang *et. al* [19]+PD	0.31	10.0 ± 6.6	5.21	0.30	10.6 ± 6.8	5.40
		Ours	**0.73**	$\mathbf{4.4 \pm 7.4}$	**3.33**	**0.71**	$\mathbf{4.9 \pm 6.6}$	**3.40**
	IMG	Grady *et. al* [12]+PD	0.67	5.1 ± 6.1	14.94	0.60	6.5 ± 5.8	14.00
		Ours	**0.88**	$\mathbf{1.4 \pm 3.4}$	**2.67**	**0.81**	$\mathbf{1.9 \pm 3.6}$	**2.08**

Table 1. **Static grasp evaluation**. We compare our model with grasp samples from the DexYCB dataset (MC), generated samples by a grasp synthesis method on the DexYCB and HO3D object sets (SYN), and samples extracted from an image-based hand pose estimator (IMG). We evaluate the baseline grasps in the simulation via PD-control (*+PD) directly or after de-noising via inverse kinematics (*+IK) for the motion capture data. We observe that our method outperforms the baselines in all metrics and conditions. The results indicate that static grasp references 1) will not lead to stable grasps when evaluated in a physics simulation and 2) suffer from interpenetration. Our method improves the interpenetration and learn stable grasps in a dynamic setting.

While the grasping policy maintains a stable grasp, the motion synthesis module takes over the control of the 6D pose of the hand. Similar to the grasping policy, we propose a feature extraction layer that incentivizes the model to move the hand to a target pose with the object in-hand.

To control the global hand motion, we estimate a 6D target pose for the hand: $\widehat{\mathbf{T}}_h = \psi(\mathbf{s}, \mathbf{T}_g, \mathbf{D})$. In particular, we estimate the global target hand pose $\widehat{\mathbf{T}}_h$ by computing the distance between the object's current 6D pose $\mathbf{T}_o$ and the target 6D pose $\overline{\mathbf{T}}_o$. We then translate and rotate the hand according to the displacement using closed-loop control. Hence, the displacement is recomputed after every action. For more details, please refer to supp. material.

4. Experiments

We conduct several experiments to analyse the performance of our method. We first introduce the data and experimental details in Sections 4.1 and 4.2. Next, we show that our method can learn stable grasps and correct imperfect labels in Section 4.3. Lastly, we evaluate the motion synthesis task and provide ablations to highlight the importance of our method's components in Sections 4.4 and 4.5.

4.1. Data

DexYCB We make use of the DexYCB dataset [5]. The dataset consists of 1000 sequences of object grasping, with 10 different subjects and 20 YCB objects [3]. We filter out all left handed sequences and create a random 75%/25% train/test-split over all sequences and subjects. The data sequence contains 6D global poses for the hand and objects in the camera frame and the local joint angles, hence providing sequences of $\left\{(\overline{\mathbf{q}}_h, \overline{\mathbf{T}}_h, \overline{\mathbf{T}}_o)\right\}_{t=1}^{T}$. The data also includes meshes for the hand and objects, and the camera parameters. We determine the grasp label based on the object's displacement with regards to its initial position. The time-step with an object displacement greater than a pre-determined threshold is chosen to be the target grasp $\mathbf{D}$. Furthermore, we use a recent state-of-the-art grasp synthesis method [19] to generate grasp labels for all the objects in DexYCB and create a 400/200 label train/test-split.

HO3D We use generated grasp labels from static grasp synthesis [19] or from an image-based pose estimator after offline optimization [12] for the HO3D objects. We create a train/test-split that is proportional to the DexYCB split, which results in a 200/100 label train/test-split.

4.2. Experimental Details

We train policies by using our implementation of the PPO algorithm [35] and run simulations in RaiSim [18]. For each sequence, we initialize the environment with an object and a grasp label. The hand is initialized with a pose from earlier steps at a pre-determined distance from the object. First, we train the grasping policy with all training labels and objects. Then we continue with the motion synthesis component given the pretrained grasping policy.
We evaluate physical plausibility of a grasp in terms of stability and interpenetration on a set of unseen grasp labels and unseen objects. We define a set of complementary metrics to quantify performance extensively.

4.2.1 Metrics

Success Rate: We define the success rate as the primary measure of physical plausibility. It is measured as the percentage of sequences which maintain a stable grasp, i.e., where the object does not slip for a period of time.

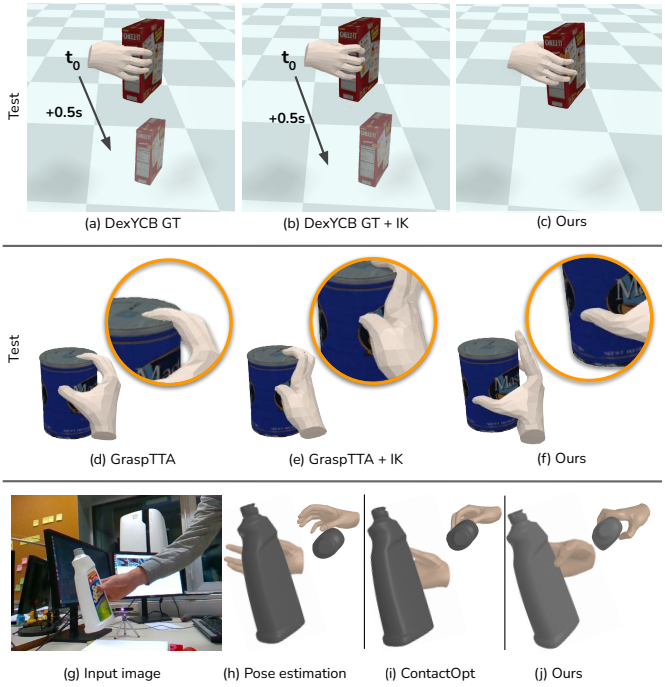

Figure 4. **Qualitative evaluation**. (a)-(c): static grasp labels often do not lead to stable grasps when evaluated in a physics simulation (a-b), which can be successfully corrected by our method (c). For an animated demonstration, please see the **video** in supp. material. (d)-(f): showcases artifacts such as interpenetration when using a state-of-the-art grasp synthesis method [19] (d-e). Our method (f) can correct such cases and generate physically-plausible grasps. (g)-(j): using images (g) to estimate an initial grasp (h). Physically implausible poses occur even with corrections via offline optimization (i), which can be corrected by our method (j).

Interpenetration: We calculate the amount of hand volume that penetrates the object. We compute it using the vertices of the MANO mesh [34] and the high-resolution object meshes.
Simulated Distance: Similar to the metric proposed in [19], we compute the mean displacement of the object. Instead of measuring the absolute displacement, we report the mean displacement in mm per second.
Contact Ratio: For the ablation study, we measure the ratio between the target contacts defined via the grasp label $\mathbf{D}$ and the contacts achieved in the physics simulation.
MPE: The mean position error between the object's position and target 3D position (for motion synthesis).
Geodesic: The angular distance between the object's current and target orientation (for motion synthesis).

4.2.2 Baselines

***+PD:** Similar to [19], we place the object into the hand via the grasp label. We then attempt to maintain the grasp using PD-control in the physics simulation.
***+IK:** We employ an offline optimization to correct for imperfections (i.e., minor distances or penetrations) in the label. The improved samples are passed to the PD-control.
Flat-RL: We employ an RL baseline that does not separate the grasping from the motion synthesis phase, but trains the full dynamic grasp synthesis task end-to-end.
Ours+static grasp: In this variant, we use our grasping policy for the grasping phase. During motion synthesis, we use PD-control to maintain the pose while the grasping policy is frozen and not actively interacting with the object.

Models	Success ↑	SimDist [mm/s] ↓	Interpenetration [cm^3]↓
GT+PD	0.30	13.7 ± 9.2	4.41
GT+IK	0.38	11.7 ± 9.4	9.08
Ours	**0.56**	$\mathbf{9.0 \pm 10.4}$	**1.74**

Table 2. **Generalization**. We evaluate generalization to unseen objects and compare our model with the baselines. We create six different test sets of three objects each, which we leave out during training. We report the average performance over all test sets.

4.3. Grasping Objects

In this experiment, we show that our method can learn to achieve stable grasps and that static grasp reference data is inherently bound to fail in a dynamic setting. We first train with labels from DexYCB [5] and further demonstrate that our approach also works with, and improves upon, labels obtained from state-of-the-art grasp synthesis method [19], on both the DexYCB and HO3D object sets. Lastly, we present results using an image-based hand pose estimator on HO3D images and labels from ContactOpt [12].

We present quantitative evaluations in Tab. 1 and qualitative results in Fig. 4. Compared to the baselines, our method is able to achieve significantly better performance on all the metrics. Importantly, the grasping policy can improve the success rate, while minimizing interpenetration (an important metric in the grasp synthesis literature). We note that our method achieves 0 interpenetration loss when evaluated in the physics simulation. In Tab. 1, however, we report interpenetration on the original MANO hand model and detailed object meshes. For computational efficiency during training, the hand model and the object meshes are simplified in the physics simulation (Section 3.1), limiting the performance of our model when evaluated in the original setting with regards to interpenetration. We found no improvement with IK for the generated (SYN) or image-based (IMG) experiments and hence omit it from the results. The improved performance in the image setup compared to other settings is due to the high-quality grasp references from [12], which already optimizes for contact. In general, there is a performance drop when moving to unseen test labels. We also find that our approach may struggle with thin

Models	MPE [mm] ↓	Geodesic [rad.] ↓
Flat-RL	0.55	1.66
Ours+static grasp	0.45	1.46
Ours+learned policy	0.30	0.92
Ours	**0.08**	**0.52**

Table 3. **Evaluation of motion synthesis**. We compare our model with a standard RL baseline (Flat-RL) and different variants of our method. We observe that our hierarchical framework outperforms Flat-RL. Furthermore, an active grasping policy during motion synthesis is key to solving the task, as indicated by the performance drop for Ours+static grasp.

objects which are difficult to grasp on a surface. For a detailed analysis and failure cases, we refer to supp. material.

Generalization to Unseen Objects To evaluate the generalization performance on unseen objects, we train and test our model on six separate train/test splits with varying complexity. Each test set consists of three objects from the DexYCB dataset. The remaining objects are used for training a policy. We average the results over all test sets and report the results in Tab. 2. While there is room for improvement in overall success rate, our method outperforms the baseline in all metrics. We provide a more detailed analysis in supp. material.

4.4. Motion Synthesis

We now demonstrate our method's ability to synthesize motions with the grasped object in hand. The goal of this task is to grasp an object and generate a trajectory that brings the object to a target 6D pose. We use a subset of representative YCB objects and create a test set with 100 randomly sampled, out-of-distribution poses $\mathbf{T}_g$. We compare against a standard RL baseline (Flat-RL) and a variant of our method that only *maintains* the pose instead of actively grasping the object (Ours+static grasp). We also compare against a learning-based motion synthesis policy (Ours+learned policy). As shown in Tab. 3, the hierarchical separation in our method is crucial for solving the task. Moreover, the decrease in performance when the hand pose is simply maintained (Ours+static grasp) solidifies the contribution of our approach. This implies that active control of the hand throughout the sequence is mandatory to maintain a stable grasp. Lastly, our method outperforms the learning-based variant (Ours+learned policy) of our motion synthesis module by a large margin on both metrics.

4.5. Ablations

In this experiment, we analyze different components of our method and show that they are crucial for achieving stable grasps. To this end, we ablate our method with different feature spaces and reward functions. We select a subset of representative objects and evaluate on our train-split of DexYCB (Section 4.1). To validate our feature extraction layer and in particular the goal space (Section 3.2.1), we compare to a variant of our approach using the original state space (w/o FeatLayer) and a variant without the goal space (w/o GoalSpace). Furthermore, we evaluate our method without the contact reward (w/o ContactRew) and without the proposed wrist-guidance (w/o WristGuidance) as proposed in Section 3.2.3. Tab. 4 shows that each component yields considerable performance improvement. We emphasize that the contact reward and a suitable feature representation are key for achieving stable grasps.

Models	Success ↑	SimDist [mm/s] ↓	Contact Ratio ↑
w/o ContactRew	0.0	24.18 ± 1.58	0.02
w/o GoalSpace	0.28	14.21 ± 10.50	0.18
w/o FeatLayer	0.47	9.69 ± 10.26	0.21
w/o WristGuidance	0.58	7.88 ± 10.57	0.28
Ours	**0.89**	$\mathbf{4.83 \pm 1.71}$	**0.43**

Table 4. **Ablations**. We ablate our proposed components. All components together comprises our method. We observe that each component increases the performance significantly in all metrics.

5. Discussion and Conclusion

In this work we have made several contributions. First, we have introduced the task of *dynamic grasp synthesis* for human-object interactions. To take a meaningful step into this direction, we leverage a physics simulation to generate sequences of hand-object interactions that are natural and physically plausible. We propose an RL-based solution that learns from a single external grasp label. We demonstrate that our method can learn stable grasps and generate motions with the object-in hand without slipping. Furthermore, we have provided evidence that our method can achieve generalization to unseen objects. While this proof of concept experiment indicates that our method works if a static hand pose reference for the unseen object is available, the method could be scaled to even larger train/test sets in the future. Finally, dynamics components such as friction of surfaces, inertia, or the center of mass are assumed to be known a priori, which is often not the case in real world settings. Adding a perceptual component to estimate these properties is a promising direction for future work.

Acknowledgements: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No 717054. Muhammed Kocabas is supported by the Max Planck ETH Center for Learning Systems. We thank Marco Bagatella, Manuel Kaufmann, Thomas Langerak, and Adrian Spurr for the fruitful discussions and help throughout this project. Lastly, we thank Alexis E. Block for the voice-over.

[1] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8709–8719, 2019. 2

[2] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 361–378. Springer, 2020. 2

[3] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *International Conference on Advanced Robotics (ICAR)*, 2015. 6

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12417–12426, 2021. 2

[5] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 4, 6, 7

[6] Tao Chen, Jie Xu, and Pulkit Agrawal. A simple method for complex in-hand manipulation. In *5th Annual Conference on Robot Learning (CoRL)*, 2021. 2

[7] Sammy Christen, Stefan Stevšić, and Otmar Hilliges. Guided deep reinforcement learning of control policies for dexterous human-robot interaction. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2161–2167, 2019. 3

[8] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5031–5041, 2020. 2

[9] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6608–6617, 2020. 2

[10] Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta. Use the force, luke! learning to predict physical forces by simulating effects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[11] Guillermo Garcia-Hernando, Edward Johns, and Tae-Kyun Kim. Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9561–9568. IEEE, 2020. 2, 3

[12] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1471–1481, 2021. 2, 6, 7

[13] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 4

[14] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 571–580, 2020. 1, 2

[15] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 1, 2

[16] Kaijen Hsiao and Tomas Lozano-Perez. Imitation learning of whole-body grasps. In *IEEE/RSJ international conference on intelligent robots and systems*, pages 5657–5662. IEEE, 2006. 3

[17] Wenlong Huang, Igor Mordatch, Pieter Abbeel, and Deepak Pathak. Generalization in dexterous manipulation via geometry-aware multi-task learning. *arXiv preprint arXiv:2111.03062*, 2021. 2

[18] Jemin Hwangbo, Joonho Lee, and Marco Hutter. Per-contact iteration method for solving contact dynamics. *IEEE Robotics and Automation Letters*, 3(2):895–902, 2018. 6

[19] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6, 7

[20] Yifeng Jiang, Michelle Guo, Jiangshan Li, Ioannis Exarchos, Jiajun Wu, and C Karen Liu. Dash: Modularized human manipulation simulation with vision and language for embodied ai. In *The ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 1–12, 2021. 3

[21] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *International Conference on 3D Vision (3DV)*, 2021. 1, 2

[22] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 2

[23] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, October 2021. 2

[24] Kailin Li, Lixin Yang, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. ArtiBoost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *arXiv preprint arXiv:2109.05488*, 2021. 1

[25] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14687–14697, 2021. 2

[26] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. In *Advances in Neural Information Processing Systems*, 2021. 3

[27] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning (CoRL)*, 2021. 3

[28] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 3

[29] Mariem Mezghanni, Malika Boulkenafed, Andre Lieutier, and Maks Ovsjanikov. Physically-aware generative network for 3d shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9330–9341, June 2021. 3

[30] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub W. Pachocki, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *CoRR*, abs/1808.00177, 2018. 2

[31] Yuxzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. *arXiv preprint arXiv:2108.05877*, 2021. 3

[32] Aravind Rajeswaran*, Vikash Kumar*, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018. 2, 3

[33] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[34] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6), Nov. 2017. 4, 7

[35] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 6

[36] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics (ToG)*, 40(4), July 2021. 3

[37] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)*, 39(6), dec 2020. 3

[38] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[39] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, 2019. 1, 2

[40] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11097–11106, 2021. 2

[41] Yuting Ye and C Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *ACM Transactions on Graphics (ToG)*, 31(4):1–10, 2012. 2

[42] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7159–7169, June 2021. 3

[43] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4), July 2021. 2, 4

[44] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15741–15751, 2021. 2